A Unified Learning Framework for Auto Face Annotation by Mining Web Facial Images

Dayong Wang, Steven C.H. Hoi, Ying He
School of Computer Engineering, Nanyang Technological University, Singapore
{s090023, chhoi, yhe}@ntu.edu.sg

ABSTRACT

Auto face annotation plays an important role in many real-world multimedia information and knowledge management systems. Recently there is a surge of research interests in mining weakly-labeled facial images on the internet to tackle this long-standing research challenge in computer vision and image understanding. In this paper, we present a novel unified learning framework for face annotation by mining weakly labeled web facial images through interdisciplinary efforts of combining sparse feature representation, content-based image retrieval, transductive learning and inductive learning techniques. In particular, we first introduce a new search-based face annotation paradigm using transductive learning, and then propose an effective inductive learning scheme for training classification-based annotators from weakly labeled facial images, and finally unify both transductive and inductive learning approaches to maximize the learning efficacy. We conduct extensive experiments on a real-world web facial image database, in which encouraging results show that the proposed unified learning scheme outperforms the state-of-the-art approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation

Keywords

web facial images, face annotation, image retrieval, sparse coding, transductive learning, inductive learning

1. INTRODUCTION

Recent years have witnessed the rapid growth of various digital and mobile devices, powerful cloud computing facilities, web 2.0 photo sharing portals and social networks. As a consequence, massive facial images have been created, distributed and shared on the internet by millions of users nowadays, in which some of the facial images are associated with tags or labels while some others are completely unlabeled. The huge amount of web facial images poses many challenges and opportunities. On the one hand, the existing huge amount of weakly labeled facial images offers an important source for knowledge discovery to tackle many long-standing research challenges, and on the other hand, the increasingly large amount of unlabeled facial images brings a critical challenge to many multimedia retrieval and knowledge management tasks. An important technique to address this challenge is *auto face annotation*, which aims to automatically assign a face with the name of the corresponding person. This technique benefits many real-world applications. For example, it can help social media portals (e.g., Facebook) to automatically annotate users' uploaded photos to facilitate the search and management of online photo albums. Besides, face annotation techniques can be applied to the news video domain where faces of key persons in a video can be automatically detected and annotated to facilitate various multimedia management tasks, such as news video summarization, retrieval and browsing [37].

Face annotation is closely related to face detection and recognition, a long-standing research challenge which has been extensively studied for years in computer vision and image processing. In general, face annotation can be formulated as a data classification problem from a machine learning and data mining perspective. It thus could be solved by two types of methodologies: "inductive learning" and "transductive learning." Below we briefly introduce some basics and existing approaches in each type of learning methodology to attack the face annotation problem.

To solve face annotation from the view of "inductive learning", one can apply some classical inductive (or model-based) face recognition/verification algorithms, which have been extensively studied in computer vision and pattern recognition for many years [2, 18, 52]. The inductive learning approaches can achieve impressive results when enough high quality labeled training data are available for building the models. However, such approach is often limited in several aspects: (i) it is usually time-consuming and expensive to collect a large amount of human-labeled training facial images, typically in a controlled environment; (ii) it is usually difficult to generalize the models when new training data or new persons are added, in which an intensive re-training process is often required; and (ii) last but not least, the annotation performance often scales poorly when the number of persons/classes is large.

To address the aforementioned limitations, some recent studies have attempted to explore the "transductive learning" approach by mining huge weakly labeled facial images freely available on the internet [43, 44]. Specifically, they build a large web facial image database by querying some existing web search engine according

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

to a celebrity name list. Given the nature of web images, these facial images are weakly labeled, i.e. their labels are often noisy and do not always correspond to the right human names. For the annotation task, given a query facial image, they first retrieve top k similar images from the weakly labeled facial image database, and then annotate the query facial image using some machine learning algorithm. In general, the above search-based face annotation (SBFA) scheme is a data-driven approach by exploring transductive learning methods to attack the face annotation task. Despite its promising performance, the SBFA approach also has some limitations. For example, it may have relatively poor generalization performance of unseen faces due to its nature of exploring only local information. Besides, it also suffers from the challenge of insufficient data, i.e., the web facial image database may not have enough weakly labeled facial images for some persons who are not popular or active on the internet.

In this work, we aim to address the above limitations of both inductive learning approach and transductive learning approach for face annotation. In particular, we propose a unified framework of Unifying Transductive and Inductive Learning (UTIL) for mining web facial images by combining the strengths of the two learning techniques to tackle the face annotation problem. From the "inductive learning" view, we propose a new Weak Label Laplacian Support Vector Machine (WL-LapSVM) algorithm for generating effective classification models from weakly labeled web facial images; from the view of "transductive learning", we apply the state-of-the-art Weak Label Regularized Local Coordinate Coding (WL-RLCC) algorithm [44] in the search-based face annotation framework; finally, we propose an entropy-based combination scheme to combine the annotation results from the two different learning schemes to maximize the learning efficacy. As a summary, the main contributions of this paper include:

- We propose a Unifying Transductive and Inductive Learning (UTIL) framework for mining large web facial images towards auto face annotation;
- We propose a Weak-Label Laplacian SVM algorithm for training effective classifiers from weakly labeled facial images, which is able to overcome the challenges of insufficient labeled data suffered by classical inductive learning methods.
- We conduct extensive experiments, in which our encouraging results show that the proposed unifying learning scheme outperforms the state-of-the-art technique for a real-world face annotation task.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents the proposed Unifying Transductive and Inductive Learning (UTIL) framework and gives the related algorithms in detail. Section 4 shows the experimental results of performance evaluation, and Section 5 concludes this paper.

2. RELATED WORK

Our work is closely related to several groups of research work.

The first group is face recognition and verification, a classic problem in computer vision and pattern recognition that has been extensively studied for many years [18, 52]. Comprehensive reviews can be found in [18, 21, 52, 11, 24]. Although they can be extended for face annotation [55], traditional face recognition techniques often suffer from a few common drawbacks. For example, they usually require high-quality facial image databases collected in well-controlled environments, which has partially motivated the recent emerging benchmark studies of unconstrained face detection and verification techniques on the facial images collected from the web, such as the LFW benchmark [6, 16, 20, 30].

The second group is related to generic image annotation techniques [17], which usually apply existing object recognition techniques to train classification models based on human-labeled training images or attempt to infer the correlation or joint probabilities between query images and annotation keywords [12, 13, 7, 17]. Given limited training data, semi-supervised learning methods have been widely used for image annotation [41, 33, 39]. Wang et al. proposed to refine the model-based annotation results with a label similarity graph by following a random walk approach [41, 32]. Similarly, Pham et al. proposed to annotate unlabeled facial images in video frames with an iterative label propagation scheme [33]. Although semi-supervised learning approaches can leverage both labeled and unlabeled data, its performance fairly depends on the amount of labeled data. It is usually time-consuming and expensive to collect enough high-quality labeled data to achieve satisfactory performance in large-scale scenarios. Recently, the search-based image annotation paradigm by mining web images has attracted more and more attention [46, 36, 41, 35]. A few studies in this area have attempted to develop efficient content-based indexing and search techniques to facilitate annotation/recognition tasks. For example, Russell et al. developed a large collection of web images with ground truth labels to facilitate object recognition tasks [36]. There are also several studies that aim to address the final annotation process by exploring effective label propagation [47, 39, 48, 42, 50]. For example, Tang et al. presented a sparse graph-based semi-supervised learning (SGSSL) approach to annotate web images [39].

The third group is face annotation on the collections of personal or family photos. Several studies have mainly focused on the annotation task on collections of personal/family photos [40, 10, 45, 1, 9], which often contain rich context clues, such as personal/family names, social context, GPS tags, timestamps, etc. In addition, the number of persons/classess is usually quite small, making such annotation tasks less challenging. These techniques usually achieve fairly impressive annotation results. Some techniques have been successfully deployed in commercial applications, e.g., Apple iPhoto [1], Google Picasa [2], Microsoft easyAlbum [10], and Facebook face auto-tagging solution [3].

The fourth group addresses face annotation by mining weakly labeled facial images on the web. A few studies consider a human name as an input query, and mainly aim to refine the text-based search results by exploiting visual consistency of facial images, which is closely related to automated image re-ranking problems. For example, Ozkan and Duygulu proposed a graph-based model for finding the densest sub-graph as the most related result [31]. Following the graph-based approach, Le and Satoh proposed a new local density score to represent the importance of each returned image [26]. Guillaumin et al. introduced a modification to incorporate the constraint that a face can only appear once in an image [14]. On the other hand, the generative approach such as the gaussian mixture model had also been adopted to the name-based search scheme and achieved comparable results [5, 14]. Recently, a discriminant approach was proposed in [15] to improve the generative approach and avoid the explicit computation in the graph-based approach. Inspired by query expansion [29], the performance of name-based scheme can be further improved by introducing the images of "friends" of the query name. Unlike these studies of filtering the text-based retrieval results, some studies have attempted to directly annotate each facial image with the names extracted

[1] http://www.apple.com/ilife/iphoto/
[2] http://picasa.google.com/
[3] http://www.facebook.com/

from its caption information. For example, Berg et al. proposed a possibility model which is combined with a clustering algorithm to estimate the relationship between facial images and the names in their captions [4]. For the facial images and the detected names in the same document (a web image and its corresponding caption), Guillaumin et al. proposed to iteratively update the assignment based on a minimum cost matching algorithm [14]. In their subsequent work [15], they further improved the annotation performance using distance metric learning techniques to achieve more discriminative feature in low dimensional space. However, limited progress has been reported on search-based face annotation (SBFA) scheme, which is fundamentally different from the previous studies of "text-based face annotation" and "caption-based face annotation." The SBFA scheme aims to solve a generic content-based face annotation problem, where a facial image is directly used as the input query. For example, Wang et al. proposed an Unsupervised Label Refinement (URL) algorithm to enhance the label matrix over the entire facial image database [43]. In their further work [44], the WLRLCC algorithm was proposed to fully exploit the top-ranking similar images of the query image via a unified optimization scheme of learning both local coordinate coding and refined labels. Besides, there is also some work for mainly addressing facial image retrieval task [50] which explores both local and global features for face retrieval and re-ranking.

Our work is fundamentally different from the previous studies on text/caption based face annotation because they aim to address the assignment between the existing facial images and their names appeared in their corresponding surrounding text, and generally do not support content-based annotation of a novel query facial image. In contrast, our work is closer to the emerging search based face annotation scheme [43, 44]. Unlike the previous transductive learning approaches, the proposed unified scheme unify both transductive and inductive learning approaches to maximize the learning efficacy.

The last group of related work is about machine learning techniques, including semi-supervised learning [56, 8, 53, 3] and multimodal fusion [22, 23]. One problem addressed in our framework is about small sample learning. It can be partially solved by Semi-supervised learning (SSL) techniques which have been extensively studied for several years. Among many existing approaches, Laplacian Support Vector Machines (LapSVM) [3] is one of state-of-the-art techniques. To reduce the computational complexity of LapSVM, Melacci et al. [28] focused on the primary Laplacian Support Vector Machines problem and proposed an efficient solution with preconditioned conjugate gradient. Weighted Margin Support Vector Machines (WMSVM) [49] is another way to solve the small-sampling problem by generalizing the original Support Vector Machines for incorporating prior knowledge. The fuzzy membership is introduced in the fuzzy support vector machine [27] such that different input points can make different contributions to the learning of decision surface. In our framework, as the number of positive samples is rarely small (only 1) and all the facial images are assigned with weak name information, motivated by the similar methodology, we proposed a new Weak Label Laplacian Support Vector Machine (WL-LapSVM) algorithm for generating effective classification models from weakly labeled web facial images.

3. UNIFIED SCHEME OF MINING WEB FACIAL IMAGES FOR FACE ANNOTATION

3.1 Framework and Overview

In this section, we briefly introduce the proposed framework of Unifying Transuctive and Inductive Learning (UTIL) for auto face annotation. It combines both transductive and inductive learning techniques in a systematic approach. Figure 1 illustrates the system flow of the proposed framework, which consists of the following three stages: (1) Preprocess the query facial image, including face detection, face alignment and facial feature extraction; (2) Apply "transductive learning" and "inductive learning" respectively on the weakly-labeled face image database; and (3) Combine the annotation results from the "transductive learning" and "inductive learning" steps, and output the final annotation. The details of each stage are described as follows.

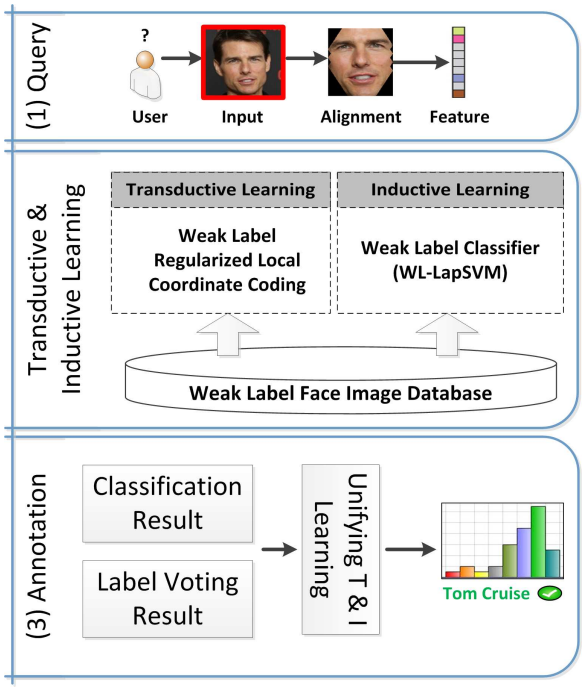

Figure 1: The Unifying Transductive and Inductive Learning (UTIL) framework for auto face annotation problem.

The first stage, as shown in Figure 1(1), is to pre-process a query facial image, including face detection, face alignment, and facial feature representation. In particular, for facial region detection and alignment, we adopt the unsupervised face alignment technique (DLK) in [54] which attempts to align all the facial images into a consistent position. We extract the GIST features [38] as the facial representation. According to our empirical study, for the aligned facial image achieved by DLK algorithm, the GIST feature performs better than the other facial features (e.g. Gabor, color, edge, or raw image intensity). As a result, each face is represented with a 512-dimensional vector in our framework.

The second stage, as shown in Figure 1(2), consists of two independent learning steps: (i) annotation by "transuctive learning" and (ii) annotation by "inductive learning." Both are applied on the same web facial image database. To build such a large-scale facial image database, we can choose a list of desired human names and submit them to some existing web search engine (e.g., Google in our approach) for crawling their related web facial images. As the output of this crawling process, we obtain a collection of web facial images, each of them is associated with a human name. Given the nature of web images and the limitation of search engine, these facial images are usually noisy, and the name labels may be incorrect or incomplete, especially for the less popular persons. We thus refer to such web facial images with noisy names as weakly

labeled facial images. For each image in weakly labeled facial image database, the same pre-processing step as the previous step is applied and no-face-detected images are removed.

For the "transductive learning" step, we apply the state-of-the-art Weak Label Regularized Local Coordinate Coding (WLRLCC) algorithm in a search based face annotation paradigm [44], which aims to annotate the query image by fully exploring the top-n similar images and their corresponding labels. For this problem, two key factors affect its final annotation performance: (1) Generating more represented feature for re-ranking as all the top ranking images are close to each other in the original feature space; (2) Enhancing the initial weak labels. In WLRLCC algorithm, these two problems are tackled simultaneously in one optimization problem.

For the "inductive learning" step, we have to address the problem of insufficient labeled data for training effective classifiers. Since the number of images and the number of persons are both large in the web facial image database, it is impossible and impractical to label all the facial images due to the expensive human labeling costs. In our framework, we assume that only one facial image can be manually labeled for each person. On the other hand, all the facial images are weakly labeled during the crawling step. As a result, the core problem is how to effectively train classifiers based on a small number of well labeled data and a large amount of weakly labeled data. To tackle issue, a natural choice is to explore semi-supervised learning techniques, e.g., semi-supervised support vector machines. However, the conventional semi-supervised learning techniques cannot deal with weakly labeled data properly. In this paper, we propose the Weak Label Laplacian Support Vector Machines (WL-LapSVM) algorithm to overcome the challenge.

The third step is about the combination of the annotation results of the previous trasductive and inductive learning stages. To this purpose, we evaluate several last fusion scheme to merge the two annotation results. We also proposed an entropy based weighting combination scheme, which achieve fairly good fusion result with less computation effort.

3.2 Transductive Learning via WLRLCC

In this section, we briefly introduce the search-based face annotation (SBFA) scheme and the Weak Label Regularized Label Local Coordinate Coding (WLRLCC) algorithm, which are proposed in [44] and employed in the "transductive learning" step of our UTIL framework.

3.2.1 Preliminaries

Throughout the paper, we denote the matrixes by upper case letters, e.g. X, D; we denote the vectors by bold lower case letters, e.g. $\mathbf{x}, \mathbf{x}_i$; we denote the scalars by the normal letters, e.g. x_i, x_{ij}, X_{ij}, where x_i is the i-th element of the vector $\mathbf{x}$, x_{ij} is the j-th element of the vector $\mathbf{x}_i$, and X_{ij} is the element in the i-row and j-column of the matrix X.

3.2.2 Weak Label Regularized Local Coordinate Coding

For the SBFA scheme, consider a query facial image $\mathbf{x}_q \in \mathbb{R}^d$ in a d-dimensional feature space, we firstly retrieve its top n similar images $X = \{(\mathbf{x}_i, \mathbf{y}_i)_{i=1,2,\ldots,n}\}$ from the weakly labeled facial image database, where $\mathbf{y}_i \in \{0,1\}^m$ is the name label vector of its corresponding facial image $\mathbf{x}_i$, $\|\mathbf{y}_i\|_0 = 1$, and m is the total number of classes (names) among all the top-n facial images. For the annotation task, one baseline algorithm is to adopts a soft-max weighted majority voting scheme with these initial label information $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, which is refereed as "SMW" in the following sections. The "SMW" method is limited in two aspects, 1). the initial label information is noisy; 2). the retrieval results can be refined in more powerful feature representation space. The WLRLCC algorithm address these two problems in a unify framework with two iterative steps: the *Coding Learning* step and the *Label Learning* step.

The purpose of *Coding Learning* is to obtain a more discriminative local coordinate coding representation, where the local coordinate coding technique is adopted [51]. For the i-th facial image $\mathbf{x}_i$, its sparse representation $\mathbf{s}_i$ is reconstructed by solving the problem $e(\hat{\mathbf{s}}_i; \mathbf{x}_i)$ based on the dictionary $B = [X, I] \in \mathbb{R}^{d\times(n+d)}$, where $X \in \mathbb{R}^{d\times n}$ is the feature matrix of the top-n similar facial images and I is an identity matrix:

$$e(\hat{\mathbf{s}}_i; \mathbf{x}_i) = \min_{\hat{\mathbf{s}}_i} \frac{1}{2}\|\mathbf{x}_i - B\hat{\mathbf{s}}_i\|^2 + \lambda \sum_{k=1}^{n+d} \hat{s}_{ik}\|B_{\star k} - \mathbf{x}_i\|^2 \quad \text{s.t. } \hat{s}_{ii} = 0 \text{ and } \hat{s}_{ij} \geq 0, j = 1, 2, \ldots, n+d \tag{1}$$

where $\mathbf{s}_i$ is a sub-vector of $\hat{\mathbf{s}}_i$ with its top-n element: $\hat{\mathbf{s}}_i = [\mathbf{s}_i, \xi_i]$, ξ_i is related to the noise information, λ is the parameter for the locality constraints, and $B_{\star k}$ is the k-th column of dictionary B. As a result, the whole formulation for all the top-n facial images is as follows

$$E_1(\hat{S}; X) = \sum_{i=1}^{n} e(\hat{\mathbf{s}}_i; \mathbf{x}_i) \tag{2}$$

where $\hat{S} \in \mathbb{R}^{(n+d)\times n} = [S; \Xi]$, $S \in \mathbb{R}^{n\times n}$ is the non-negative local coordinate coding of X, and $\Xi \in \mathbb{R}^{d\times n}$ is the noise matrix.

The purpose of *Label Learning* is to refine the initial weak label information. The new label matrix is achieved based on the graph-based label smoothness principle, which means that two similar facial images tend to share the similar labels. In the Weak Label Regularized Label Local Coordinate Coding algorithm, the visual similarity information is introduced with the locality coding representation achieved in the previous step. In particular, the j-th local coefficient s_{ij} of facial image $\mathbf{x}_i$ essentially encodes the locality information between $\mathbf{x}_i$ and $\mathbf{x}_j, j \neq i$. A larger value of s_{ij} indicates that $\mathbf{x}_j$ is more representative of $\mathbf{x}_i$, as a result, a larger value of s_{ij} implies that the name labels of $\mathbf{x}_i$ and $\mathbf{x}_j$ are more likely to be the same. Suppose the initial weak label matrix is $\tilde{Y}$, the objective function for the refined label matrix Y is as follows:

$$E_2(Y; S) = \min_{Y\geq 0} \frac{1}{2}\sum_{i,j} s_{ij}\|Y_{i\star} - Y_{j\star}\|^2 + \lambda\|(Y - \tilde{Y}) \circ M\|_{\mathrm{F}}^2 \tag{3}$$

where $M = [h(\tilde{Y}_{ij})]$ is an indicator matrix: $h(x) = 1$ if $x > 0$ and otherwise $h(x) = 0$, and $\circ$ denotes the Hadamard product of two matrices. s_{ij} is the j-th local coefficient of facial image $\mathbf{x}_i$, which essentially encodes the locality information between $\mathbf{x}_i$ and $\mathbf{x}_j, j \neq i$. As the ideal true label matrix is often very sparse, a series of extra convex sparsity constraint are introduced to take into the consideration of sparsity: $\|Y_{i\star}\|_1 \leq 1$, where $i = 1, 2, \ldots, n$

To better exploit the potential of the two previous learning approaches:*Code Learning* and *Label Learning*, they are further reinforced into a unified optimization framework. Specifically, the optimization formulation of Weak Label Regularized Label Local Coordinate Coding is formulated as follows:

$$\begin{aligned} &Q(\hat{S}, Y) = E_1(\hat{S}; X) + E_2(Y; S) = \min_{\hat{S}, Y} \frac{1}{2}\|B\hat{S} - X\|_{\mathrm{F}}^2 + \\ &\lambda_1 tr(\mathbf{1}\cdot(\hat{S} \circ V)) + \lambda_2 tr(Y^\top L Y) + \lambda_3\|(Y - \tilde{Y}) \circ M\|_{\mathrm{F}}^2 \\ &\text{s.t. } \hat{S}_{ii} = 0, \|Y_{i\star}\|_1 \leq 1, i = 1, 2, \ldots, n,\ \hat{S} \geq 0, Y \geq 0 \end{aligned} \tag{4}$$

where $V \in \mathbb{R}^{(n+d)\times n}$, $V_{ij} = \|B_{\star i} - X_{\star j}\|^2$, $L = D - S$, D

is a diagonal matrix, with $D_{ii} = \frac{\sum S_{i*} + \sum S_{*i}}{2}$, $Y \in \mathbb{R}^{n \times m}$, $\mathbf{1}$ is all-one-element matrix with dimension $n \times (n+d)$, and $tr(\cdot)$ denotes a trace function. In the above, $\lambda_2 tr(Y^\top LY)$ is a label smoothness regularizer which connects the label matrix and the sparse features. For the final annotation step, an effective sparse reconstruction scheme is applied.

The reasons that we adopt the WLRLCC algorithm for the "transductive learning" step in the proposed UTIL framework are twofold: (i) the WLRLCC algorithm is suitable for handling large-scale problem as it is applied only to the short list of the similar images for each query image and is independent of the entire retrieval database size; (ii) the WLRLCC algorithm fully exploits the short list of top-ranking similar images via a unifying optimization scheme and achieves the best annotation performance over a large-scale web facial image database.

3.3 Inductive Learning via WL-LapSVM

In this section, we present the proposed Weak Label Laplacian SVM (WL-LapSVM) algorithm for solving the "inductive learning" task in the proposed Unifying Transductive and Inductive Learning (UTIL) scheme.

3.3.1 Preliminaries

We denote the whole retrieval database and the corresponding name labels with $D = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 1, 2, \ldots, \bar{n}\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional facial feature vector, $\mathbf{y}_i \in \{0,1\}^{\bar{m}}$ is the corresponding name label vector with only one non-zero value:$\|\mathbf{y}_i\|_0 = 1$, $\bar{n}$ denotes the total number of facial images in the whole database, and $\bar{m}$ is the total number of unique human names. For simplicity, in the following sections, we denote the label vector $\mathbf{y}_i$ by y_i, which equals the index value of the non-zero item in $\mathbf{y}_i$. Further, we denote by $D^j = \{(\mathbf{x}_k, y_k) \in D | y_k = j\}$ the subset of the crawled images belonging to the j-th name(person).

In order to train inductive classifiers for each person, we manually label a small number of facial images as the preliminary set of labeled images. As a result, for the j-th name(person) in the retrieval database, its image set D^j can be further divided into two subsets: the *label set* L^j and the *unlabel set* U^j, where $D^j = L^j \bigcup U^j$. In particular, in our experiment only one image is labeled for each name (person), which means $|L^j| = 1, j = 1, 2, \ldots, \bar{m}$. This kind of setting is reasonable, since in real-world application the number of names is very large and it is time-consuming and impractical to label a large amount of labeled data.

As there is only one positive sample for each class(person), general semi-supervised learning technique, e.g. the Laplacian SVM, can be used to solve the small-sample problem. However, it does not work well in our problem due to the limited number of positive samples. We also notice all the facial images in the database are assigned with weak labels, which can be employed for prior information for classifier learning. To address the previous problem, we propose a variant of Laplacian SVM algorithm, the Weak Label Laplacian Support Vector Machine (WL-LapSVM), to handle the noisy web images in our weakly labeled facial image database.

3.3.2 Problem Formulation of WL-LapSVM

For each class(person), we will train a separate classification model $\mathbf{f}^j, j = 1, 2, \ldots, \bar{m}$, which determine whether a facial image $\mathbf{x}$ belongs to class j or not. In particular, if $\text{sign}(\mathbf{f}^j(\mathbf{x}) > 0$, then the facial image $\mathbf{x}$ is supposed to be in the j-th class and be labeled with the j-th name.

Generally, for the j-class(person), we can construct its training set T^j by introducing the labeled subset L^j from the j-th class as the labeled positive samples, introducing the labeled subsets $\{L^k | k \neq j\}$ from the other classes as the labeled negative samples, and the unlabeled subset U^j from the j-th class as the unlabeled samples. However, in our application, the size of labeled positive training set L^j is extremely small(only 1), as a result, we only collect a subset of the labeled negative samples into the training sets T^j. For simplicity, we can represent the training set T^j of the j-th class as follows:

$$T^j = \{(\mathbf{x}_1, z_1), (\mathbf{x}_2, z_2), \ldots, (\mathbf{x}_l, z_l), \mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \ldots, \mathbf{x}_{l+u}\}$$

where $z \in \{-1, 1\}$ is the class labels, l is the number of labeled samples, and u is the number of unlabeled samples ($l + u = |T^j|$). Following the traditional Laplacian SVM algorithm (LapSVM) [28], we define a kernelized target function $\mathbf{f}^j(\mathbf{x})$ for the j-class, which is also denoted as $\mathbf{f}(\mathbf{x})$ for short in the rest parts of this section:

$$\mathbf{f}(\mathbf{x}) = \sum_{k=1}^{l+u} \alpha_k \kappa(\mathbf{x}_k, \mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{k}(\mathbf{x}), \mathbf{x}_k \in T^j \tag{5}$$

where $\kappa(\cdot, \cdot)$ is a kernel function, $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_{l+u}]$ and $\mathbf{k}(\mathbf{x}) = [\kappa(\mathbf{x}_1, \mathbf{x}), \kappa(\mathbf{x}_2, \mathbf{x}), \ldots, \kappa(\mathbf{x}_{l+u}, \mathbf{x})]$. The traditional Laplacian SVM problem is to minimize the following objective function with respect to the previous classification function $\mathbf{f}$.

$$\min_{\mathbf{f}(\mathbf{x})} g(\alpha) = \sum_{k=1}^{l} V(\mathbf{x}_k, z_k, \mathbf{f}) + \frac{\lambda_1}{2}\Phi(\mathbf{f}) + \frac{\lambda_2}{2}\|\mathbf{f}\|_A \tag{6}$$

where $\|\cdot\|_A$ is the norm in the Reproducing Kernel Hilber Space (RKHS) $\mathcal{H}_k$ of kernel κ. V is a loss function on the label data, we choose the L_2 hingloss function in our experiment :

$$V(\mathbf{x}_k, z_k, \mathbf{f}) = \frac{1}{2}\sum_{k=1}^{l} \max(1 - z_k \mathbf{f}(\mathbf{x}_k), 0)^2 \tag{7}$$

$\Phi(\mathbf{f})$ is an intrinsic regularizer to employ the geometry information among the label data and unlabelled data by the Laplacian matrix, which is defined as :

$$\Phi(\mathbf{f}) = \sum_{\mathbf{x}_i, \mathbf{x}_j \in T^j} W_{ij}(\mathbf{f}(\mathbf{x}_i) - \mathbf{f}(\mathbf{x}_j))^2 = \boldsymbol{\alpha}^\top KLK\boldsymbol{\alpha} \tag{8}$$

where K is the kernel matrix of the instances in the training set T^j, L is the graph Laplacian matrix, and $L = D - W$, W is the adjacency matrix of the data graph in training set ($W_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j), \mathbf{x}_i, \mathbf{x}_j \in T^j$, we choose κ as an RBF kernel function in our framework). D is a diagnal matrix with diagonal elements as $D_{ii} = \sum_j W_{ij}$.

In our application, the traditional Laplacian SVM algorithm does not work well, because the positive labeled images are very limited (only one positive reference image) and the facial images are in a high dimensional space and can not be separated linearly. In order to overcome this challenge, we propose to employ $p \leq u$ samples from the unlabeled samples of T^j as the pseudo-positive labeled samples, which means that they are not definitely positive samples. To reduce the risk of using these unlabeled samples, we propose to assign each unlabeled instance with a confidence weighting value, which could be achieved with extra information(e.g. the ranking position in the searching result).

As a result, the previous unlabeled samples $\{\mathbf{x}_k | k = l+1, l+2, \ldots, l+u\}$ in T^j can be represented as $\{(\mathbf{x}_k, z_k, \varepsilon_k) | k = l+1, l+2, \ldots, l+u\}$, where the confidence weighting value $\varepsilon_k \in [0, 1]$. The label z_k is set as 1 for the collected pseudo samples, and for the uncollected samples, we can just set its confidence value as $z_k = 0$. Correspondingly, for the labeled instances in T^j, we can also assign a fixed confidence weights $\varepsilon_k = 1$ for $k = 1, 2, \ldots, l$.

By employing the weighted unlabeled instances into the traditional Laplacian SVM as an extra term, we achieve a new confidence value weighted Laplacian SVM formulation as follows:

$$\min_{\mathbf{f}(\mathbf{x})} g(\alpha) = \sum_{k=1}^{l} \varepsilon_k V(\mathbf{x}_k, z_k, \mathbf{f}) + \frac{\lambda_1}{2}\Phi(\mathbf{f}) + \frac{\lambda_2}{2}\|\mathbf{f}\|_A + \lambda_3 \sum_{k=l+1}^{l+u} \varepsilon_k V(\mathbf{x}_k, z_k, \mathbf{f}) \tag{9}$$

where $\varepsilon_k = 1$ for $k = 1, 2, \ldots, l$, and $\varepsilon_k \in [0, 1]$ for $k = l+1, l+2, \ldots, l+u$. We refer to the proposed modified Laplacian SVM based on weak label as "**WL-LapSVM**" for short. If $\lambda_3 = 0$, the formulation reduces to a standard Laplacian SVM algorithm.

3.3.3 *Optimization Algorithm*

In this section, we briefly introduce the optimization algorithm for the proposed WL-LapSVM, which could be reformulated as follows:

$$\min_{\mathbf{f}(\mathbf{x})} g(\boldsymbol{\alpha}) = \frac{1}{2}\sum_{i=1}^{l+u} \varepsilon_i' \max\left(1 - z_i \mathbf{k}_i^\top \boldsymbol{\alpha}, 0\right)^2 + \frac{\lambda_1}{2}\boldsymbol{\alpha}^\top KLK\boldsymbol{\alpha} + \frac{\lambda_2}{2}\boldsymbol{\alpha}^\top K\boldsymbol{\alpha} \tag{10}$$

where $\mathbf{k}_i^\top = \mathbf{k}(\mathbf{x}_i)$, $\varepsilon_i' = 1$ with $i = 1, 2, \ldots, l$, and $\varepsilon_i' = \lambda_3 \varepsilon_i \in [0, \lambda_3]$ with $i = l+1, l+2, \ldots, l+u$. In order to solve this problem, we follow the Newton Method proposed in [28]. In each Newton's step, we update $\boldsymbol{\alpha}$ with the following rule :

$$\boldsymbol{\alpha}^{(t)} = \boldsymbol{\alpha}^{(t-1)} - sH^{-1}\nabla_{\boldsymbol{\alpha}} \tag{11}$$

where t is the iteration number, s is the step size, and $\nabla_{\boldsymbol{\alpha}}$ and H are the gradient vector and Hession matrix for $g(\boldsymbol{\alpha})$ in Equation 7. For ∇, we have:

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}} &= \sum_i \varepsilon_i' \mathbf{k}_i z_i (z_i \mathbf{k}_i^\top \boldsymbol{\alpha} - 1) + \lambda_1 KLK\boldsymbol{\alpha} + \lambda_2 K\boldsymbol{\alpha} \\ &= KSK\boldsymbol{\alpha} - KS\mathbf{z} + \lambda_1 KLK\boldsymbol{\alpha} + \lambda_2 K\boldsymbol{\alpha}\end{aligned} \tag{12}$$

where $S \in \mathbf{R}^{n \times n}$ is a diagonal matrix and its i-th element in the main diagonal is ε_i'. K is the kernel matrix of the instances in the corresponding training set. The Hessian matrix H could be achieved as follows:

$$H = \nabla_{\boldsymbol{\alpha}}^2 = KSK + \lambda_1 KLK + \lambda_2 K \tag{13}$$

The step size s could be fixed to 1 or optimized by line searching. The iterative update is guaranteed to converge when the error vector does not change for two consecutive iterations.

3.3.4 *Name Annotation*

For each name (person) in the retrieval database, we can build a WL-LapSVM model $\mathbf{f}^j, j = 1, 2, \ldots, \bar{m}$, respectively. To annotate the query image $\mathbf{x}_q$, we firstly compute its prediction value $y_{qj} = \mathbf{f}^j(\mathbf{x}_q), j = 1, 2, \ldots, \bar{m}$; then convert these prediction value into the probability scale: $p_{qj} = \frac{1}{1+\exp(-y_{qj})}$ by fitting a sigmoid function following the technique in [34]. Finally, the annotated name list is obtained by sorting the previous probability value $\{p_{q1}, p_{q1}, \ldots, p_{q\bar{m}}\}$.

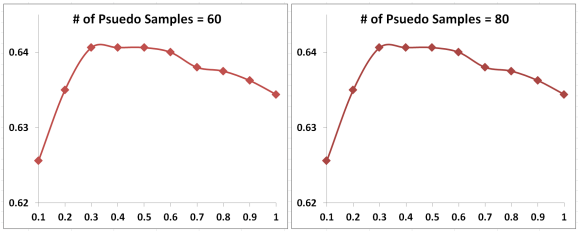

Figure 2: Annotation performance of WL-LapSVM with different confidence weighting settings. The x-axis is the confidence weight ε_p of the last pseudo sample.

3.3.5 *Pseudo Set Construction and Confidence Weights Setting*

For the proposed WL-LapSVM algorithm, there are mainly two critical problems that highly affect its annotation performance: 1) one problem is how to collect the p pseudo positive samples from the unlabeled training samples; 2) another problem is how to set the confidence value for each pseudo positive sample. In the following, we will take the j-th class as an example, where the image set is $D^j = L^j \bigcup U^j$ and the training set is T^j.

For the first problem, in our framework there are two ways to introduce the pseudo samples from the unlabeled set U^j: one way is to collect the images that are close to the true positive sample in L^j; another way is to collect the images with high Google ranking values, which means that these images are at the top-ranking position in Google retrieval result. According to our experiments, we found that the performance of the second scheme is much better than the first scheme. For example, supposed $p = 80$ extra pseudo samples are collected, the annotation performance of the second scheme is 68%, compared with 26% of the first method.

For the second problem, it aims to introduce the prior knowledge into WL-LapSVM by setting the confidence weighting value ε in Eq. 9. Suppose the p pseudo positive samples are $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\}$, sorted by the Google ranking value. In our framework, we set the confidence weight of $\mathbf{x}_i$ according to its index value i by using a monotonic decreasing function: $\varepsilon_i = \exp(\frac{i-1}{\theta})$, where θ is a parameter. By choosing different θ value, we can control the contribution of the pseudo positive sample set. In figure 2, for different number of pseudo positive samples ($p = 60$ and $p = 80$), we evaluate different confidence weights setting by making $\varepsilon_p = \exp(\frac{p-1}{\theta})$ equal $0.1, 0.2, \ldots, 1$. Obviously, when $\varepsilon_p \in [0.3, 0.6]$, the annotation performances are consistently better. As a result, we set $\varepsilon_p = 0.5$ for all the following experiments.

3.4 Fusion Strategies for Combination

The last step for the proposed UTIL framework is to combine the annotation results from both "transductive learning" and "inductive learning." Generally, the annotation problem can also be formulated as a multi-class classification problem, and the combination is a typical late-fusion (post-classification fusion) problem. There are numerous score late fusion methods in the literature [22, 23], which can be classified into three levels: *Abstract Level*, *Rank Level*, and *Measurement level*. As there are only two "classifiers" in the UTIL framework, the voting-based *abstract level* fusion is unsuitable for our experiments. For the *rank level* fusion, each model (the "transductive learning" model and the "inductive learning" model) outputs a list of possible names for the query image, sorted in decreasing order of confidence. For the *measurement level* fusion, each model outputs the possibility (confidence) values of assigning different name to the query image.

In paretical, for a query image, the annotation results of the "transductive" and "inductive" learning steps can be represented as two possibility(measurement) score vectors: $\mathbf{p}_T, \mathbf{p}_I \in [0,1]^{\bar{m}}$, respectively. For example, p_{Ti} illustrates the possibility of assigning the i-th name to the query image according for the "transductive learning" model. We can easily generate the ranking results for the two models by sorting the confidence score vector $\mathbf{p}_T, \mathbf{p}_I \in [0,1]^{\bar{m}}$, respectively.

3.4.1 Measurement Level Fusion

For the *measurement level* fusion, following the normalization scheme in [22], we use *sum rule* for measurement combination and adopt two kinds of normalization methods: the min-max normalization which is refereed as "MLF-MinMax" in the following experiments and the Z-score normalization which is referred as "MLF-Zscore" for short.

3.4.2 Rank Level Fusion

For the *rank level* fusion scheme, Ho et al. [19] describe three methods to combine the ranks assigned by the different models: *highest rank* method, *Borda count* method, and *Logistic Regression* method. In our experiments, we only adopt the borda count scheme for fusion as there are just two models in the UTIL framework. Instead of directly using the rank position as the rank value, we set the rank value according to a monotonic decreasing function. For the combination weights of different ranks, we propose two kinds of methods: one is based on confidence value regression ("RLF-Regression"), and another is based on the confidence value entropy information("RLF-Entropy").

For RLF-Regression, given a set of query images as the training set, we set the weight value w_T for the rank of the "transductive" model as 1 if the "transductive" model achieves a better annotation performance than the "inductive" model, otherwise, $w_T = 0$. Then we adopt the SVM algorithm to train a regression model for the weight value w_T by using the possibility(measurement) scores as the feature vector. Finally, for a test query image, we use the learned regression model to predict the weighting value w_T for the rank result of the "transductive" model and generate the weighting value w_I for the rank result of the "inductive" model by $1 - w_T$.

For RLF-Entropy method, we aim to avoid the computation effort in the previous regression scheme and estimate the weighting value w_T according to the entropy information of the confidence(measurement) vector. In particular, we define the entropy of the confidence vector $\mathbf{p}_T$ as:

$$\pi_T = -\sum_i \frac{p_{Ti}}{\|\mathbf{p}_T\|_1} log \frac{p_{Ti}}{\|\mathbf{p}_T\|_1}$$

Similarly, we can achieve the entropy value π_I for $\mathbf{p}_I$. It is not difficult to find that when the entropy value π_T is large, the difference of the measurement scores among different candidate names is small, which indicates that the corresponding "transductive" model is less confidence, so that we should set a small weighting value for the rank result of the "transductive" model. As a result, we set the weight w_T for the "transductive" model as:

$$w_T = 1 - \frac{\exp(\pi_T)}{\exp(\pi_T) + \exp(\pi_I)} = 1 - w_I$$

where w_I is the weight value for the rank result from the "inductive" model.

4. EXPERIMENTS

To evaluate the performance of the proposed Unifying Transductive and Inductive Learning (UTIL) scheme, we conduct an extensive set of experiments on a large real-world weakly labeled facial images database. In the following, we first briefly introduce our experiment dataset, then discuss the parameter settings, and finally present the experimental results and discussion.

4.1 Experiment Testbed

Although several web facial images databases are available, for example, LFW[4] [20], Pubfig[5] [25], Yahoo!News[6] [4, 15], and FAN-Large[7], these databases are not suitable for the performance evaluation of the proposed Unifying Transductive and Inductive Learning (UTIL) scheme for serval reasons, e.g., the number of facial images per person is too small for the search-based face annotation. In our experiments, we adopt the weakly labeled web facial image database released in [44], which consists of four retrieval database in different size and one query database with about $1,600$ images. In order to train the proposed WL-LapSVM model in the "inductive learning" step, for each person/class in the retrieval database, we manually label one front-view facial image as the reference image. Notice that we do not make extra collection for the reference image, which aims to examine the generalization performance of the proposed Unifying Transductive and Inductive Learning (UTIL) scheme. Due to the manual labeling effort, in our experiments, we only label the retrieval database "GDB-040K" which contains 400 persons and more than $40,000$ facial images.

To evaluate the annotation performance, we adopt the *hit rate* at top-T annotated results as the performance metric, which measures the likelihood of having the true label among the top-T annotated names. Specifically, for $T = 1$, the *hit rate* is the same with the *accuracy*. For the "transductive learning" step, we retrieve 40 most similar images for each query image from the retrieval database. For the "inductive learning" step, as a fair comparison, we adopt the **RBF** kernel for all the compared algorithms with $\sigma = 0.2$. For the other parameters, we randomly divide the query(test) database into two parts of equal size, and randomly collect one part for tuning the optimal parameters by a grid search scheme.

4.2 Evaluation on Positive Sample Size

In this experiment, we evaluate how the number of pseudo positive samples affects the performance of different algorithms. We compare the proposed WL-LapSVM algorithm with three baseline algorithms: the SVM algorithm using the only one positive sample, the LapSVM algorithm using the only one positive sample, and another SVM that employs p weakly labeled samples as positive samples, denoted as "WL-SVM" for short. The experimental result is presented in Figure 3 and Table 1.

We can draw several observations from the results. First, for both SVM and LapSVM, the annotation performances are rather poor by using only one positive reference image, and the hit rate of svm and LapSVM at the top-1 position are only 7.6% and 10.3%, respectively. By introducing the manifold information, the LapSVM algorithm is slightly better than SVM. Second, by using the pseudo positive samples, the annotation performance of SVM can be improved significantly. In particular, when the number of pseudo samples is increased from 10 to 150, the annotation performance will also boost from 23.0% to 62.0%. The additional pseudo positive training samples take credit for the performance improvement, which indicates it is important to employ extra samples in the "inductive learning" step. Third, the annotation performance can be further improved by adopting the proposed WL-LapSVM algorithm,

[4] http://vis-www.cs.umass.edu/lfw/
[5] http://www.cs.columbia.edu/CAVE/databases/pubfig/
[6] http://goo.gl/2XlES
[7] http://www.vision.ee.ethz.ch/~calvin/fan-large/

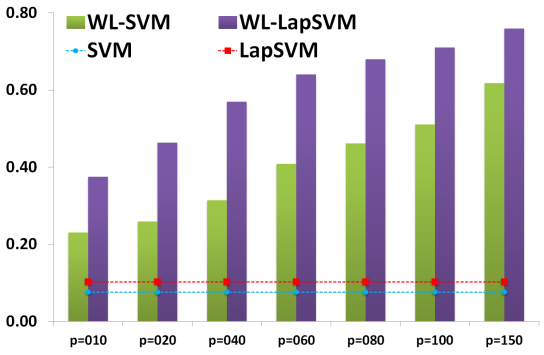

Figure 3: Annotation performance of inductive learning algorithms with varied numbers (p) of pseudo positive samples.

which indicates that the proposed WL-LapSVM algorithm can effectively use the pseudo samples by assigning different pseudo positive samples with different confidence values. For example, by using only 20 pseudo positive samples, the proposed WL-LapSVM algorithm can achieve a better annotation performance than WL-SVM with 80 pseudo positive samples. This is very important in a large-scale problem, where the number of persons is huge and more training samples will take more storage space and computational costs.

Table 1: Annotation performance of inductive learning algorithms with varied numbers (p) of pseudo positive samples

	SVM	LapSVM	WL-SVM	WL-LapSVM
$p = 10$	0.0756	0.1025	0.2300	0.3750
$p = 20$	0.0756	0.1025	0.2594	0.4631
$p = 40$	0.0756	0.1025	0.3138	0.5700
$p = 60$	0.0756	0.1025	0.4088	0.6406
$p = 80$	0.0756	0.1025	0.4613	0.6800
$p = 100$	0.0756	0.1025	0.5106	0.7106
$p = 150$	0.0756	0.1025	0.6200	0.7614

4.3 Evaluation of Auto Face Annotation

In this experiment, we evaluate the annotation performance of the "transducitve learning" step and the "inductive learning" step, respectively. For the "transductive learning" scheme, we adopt two algorithms, including a majority-voting based algorithm "SMW" and the state-of-the-art "WLRLCC" algorithm [44]. For the "inductive learning" scheme, we adopt the WL-SVM algorithm in the previous experiment and the proposed WL-LapSVM algorithm. For both WL-SVM and WL-LapSVM algorithms, we use $p = 150$ pseudo positive samples in our experiment. In Figure 4 and Table 2, both the mean and standard deviation of the annotation performance (*hit rate*) are reported with different T values, where T is the number of annotated names.

Several observations can be drawn from the above experimental results. First of all, for the "transductive learning" step, the WLRLCC algorithm significantly outperforms the simple baseline algorithm "SMW", which is similar to the observations reported in [44]. Second, for the "inductive learning" step, the proposed WL-LapSVM algorithm achieves comparable results with the state-of-the-art WLRLCC algorithm. In particular, the WL-LapSVM algorithm performs slightly worse than the WLRLCC algorithm when only one name is annotated($T = 1$), however, its performance is better for large T values. It indicates that the "inductive learning" algorithm WL-LapSVM has a better recall performance than the "transductive learning" algorithm WLRLCC.

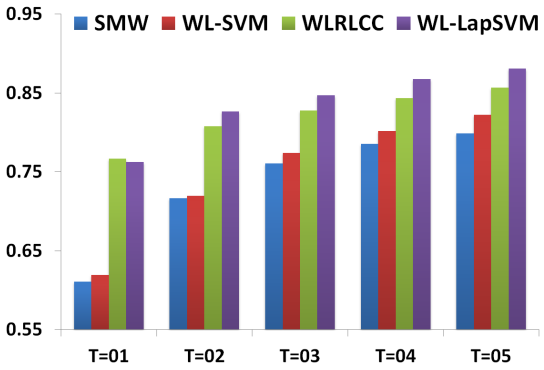

Figure 4: Comparison of face annotation performance by different algorithms, where T is the number of the annotated names.

4.4 Evaluation on Different Combinations

In this experiment, we evaluate the annotation performance of the proposed Unified Transductive and Inductive Learning (UTIL) scheme, by combining the two annotation models with different last-fusion algorithms, including the *measurement level* fusion ("MLF-MinMax", "MLF-Zscore") and the *rank level* fusion ("RLF-Regression", "RLF-Entropy"). The average annotation performance are reported in Figure 5 and Table 2, respectively.

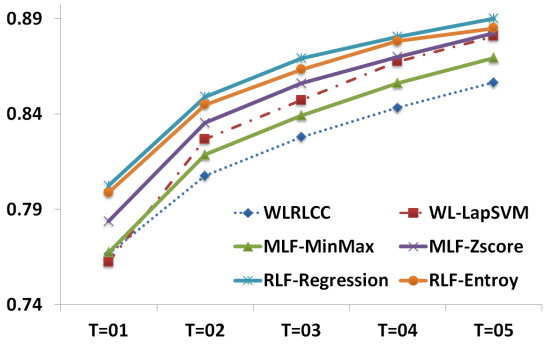

Figure 5: Comparison of different last-fusion algorithms in the UTIL framework

Several observations can be drawn from the above experimental results. First, by adopting proper fusion algorithm, the proposed UTIL scheme can significantly boost the annotation performance. In particular, the annotation results of WLRLCC and WL-LapSVM are 0.7665 and 0.7624. By using the last-fusion in UTIL, the performance can be boosted to 0.8025 by "RLF-Regression" fusion and 0.7988 by "RLF-Entropy" fusion. Second, in our experiments, the *rank level* fusion algorithms are more suitable for the proposed UTIL framework than the *measurement level* fusion algorithms. In particular, both "RLF-Regression" and "RLF-Entropy" consistently outperform the *measurement level* fusion ("MLF-MinMax" and "MLF-Zscore"). Furthermore, the "MLR-MinMax" fusion scheme even performs worse than the individual "WL-LapSVM" algorithm for large T values. Third, for the *rank level* fusion scheme, although "RLF-Regression" is slightly better, it is a supervised scheme

Table 2: Comparison of auto face annotation performance by different approaches.

T	SMW	WL-SVM	WLRLCC	WL-LapSVM	MLF-MinMax	MLF-Zscore	RLF-Regression	RLF-Entropy
01	0.6110	0.6190	**0.7665**	0.7624	0.7674	0.7838	**0.8025**	0.7988
	±0.008	±0.011	±0.013	±0.014	±0.013	±0.013	±0.012	±0.012
02	0.7168	0.7198	0.8076	**0.8268**	0.8188	0.8354	**0.8491**	0.8449
	±0.009	±0.010	±0.012	±0.012	±0.012	±0.011	±0.008	±0.008
03	0.7608	0.7739	0.8279	**0.8473**	0.8393	0.8561	**0.8693**	0.8634
	±0.008	±0.008	±0.009	±0.011	±0.009	±0.010	±0.006	±0.007
04	0.7855	0.8015	0.8433	**0.8675**	0.8563	0.8700	**0.8805**	0.8783
	±0.009	±0.008	±0.009	±0.011	±0.008	±0.007	±0.007	±0.005
05	0.7988	0.8223	0.8566	**0.8808**	0.8695	0.8824	**0.8901**	0.8850
	±0.008	±0.007	±0.008	±0.010	±0.007	±0.007	±0.010	±0.005

that needs extra labeled samples and training efforts. The proposed entropy based fusion method "RLF-Entropy" can achieve a very close combination result without extra efforts.

4.5 Evaluation on Parameter Sensitivity

For the proposed WL-LapSVM algorithm, the parameter λ_3 in Equation 9 is fixed as 1, and the parameters λ_1 and λ_2 are found by a grid search scheme. Figure 6 shows one the grid search result, where the ranges for λ_1, λ_2 are $\{0.01, 0.05, 0.1, 0.5, 1\}$. We notice that the performance of WL-LapSVM tends to be stable in the region $\lambda_1 \in [0.01, 0.2]$ and $\lambda_2 \in [0.08, 0.7]$. For this grid searching result, we choose $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$ for the further experiments. In our experiments, we also found that generally the WL-LapSVM algorithm performs well with the parameters located in the previous range, which indicates that the WL-LapSVM algorithm is robust in terms of the parameter setting.

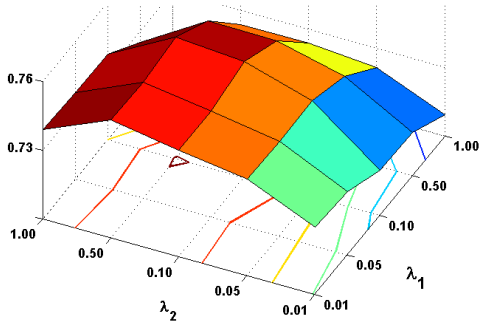

Figure 6: Grid search result of WL-LapSVM

5. CONCLUSION

This paper investigates a unifying learning scheme by combining both transductive and inductive learning techniques to mine web facial images for auto face annotation. In particular, to address the small positive sample problem in the "inductive learning" scheme, we propose a Weakly Label Laplacian Support Vector Machines (WL-LapSVM) algorithm to train classifiers based on weakly labeled data. We adopt the state-of-the-art technique WLRLCC algorithm for the "tranductive learning" scheme. To fully exploit the two types of learning paradigms, we evaluate different last-fusion algorithms on both *measurement level* and *rank level*. We also propose an entropy-based *rank level* fusion algorithm, which performs as well as the supervised regression-based fusion algorithm without extra training efforts. Our empirical results show that the proposed UTIL scheme can significantly outperform both the transductive and inductive annotation approaches. Future work explores the applications of our techniques to solve other real-world problems.

Acknowledgement

This work is supported by Singapore MOE Academic tier-1 grant (RG33/11) and Microsoft Research grant

6. REFERENCES

[1] D. Anguelov, K. chih Lee, S. B. Göktürk, and B. Sumengen. Contextual identity recognition in personal photo albums. In *CVPR*, 2007.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Tran. PAMI*, 19(7):711–720, 1997.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, 2006.

[4] T. L. Berg, A. C. Berg, J. Edwards, and D. Forsyth. Who's in the picture. In *NIPS*, 2005.

[5] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. G. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *CVPR*, pages 848–854, 2004.

[6] Z. Cao, Q. Yin, X. Tang, and J. Sun. Face recognition with learning-based descriptor. In *CVPR*, 2010.

[7] G. Carneiro, A. B. Chan, P. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Tran. PAMI*, pages 394–410, 2006.

[8] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*, MIT Press, 2006.

[9] J. Y. Choi, W. D. Neve, K. N. Plataniotis, and Y. M. Ro. Collaborative face recognition for improved face annotation in personal photo collections shared on online social networks. *IEEE Transactions on Multimedia*, 13, 2011.

[10] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *CHI*, pages 367–376, 2007.

[11] K. Delac and M. Grgic. *Face Recognition*. IN-TECH, 2007.

[12] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[13] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *ACM Multimedia*, pages 540–547, 2004.

[14] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid.

Automatic Face Naming with Caption-based Supervision. In *CVPR*, pages 1–8, 2008.

[15] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Face recognition from caption-based supervision. *International Journal of Computer Vision*, 2011.

[16] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? Metric learning approaches for face identification. In *International Conference on Computer Vision*, Sept. 2009.

[17] A. Hanbury. A survey of methods for image annotation. *J. Vis. Lang. Comput.*, 19:617–627, October 2008.

[18] E. Hjelmås and B. K. Low. Face detection: A survey. *Computer Vision and Image Understanding*, 2001.

[19] T. K. Ho, J. J. Hull, and S. N. Srihari. Decision combination in multiple classifier systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(1):66–75, Jan. 1994.

[20] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[21] R. Jafri and H. R. Arabnia. A survey of face recognition techniques. *JIPS*, 5:41–68, 2009.

[22] A. K. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2270–2285, 2005.

[23] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):226–239, Mar. 1998.

[24] M. G. Kresimir Delac and M. S. Bartlett. *Recent Advances in Face Recognition*. I-Tech Education and Publishing, 2008.

[25] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *International Conference on Computer Vision*, Oct 2009.

[26] D.-D. Le and S. Satoh. Unsupervised face annotation by mining the web. In *ICDM*, pages 383–392, 2008.

[27] C.-F. Lin and S.-D. Wang. Fuzzy support vector machines. *IEEE Transactions on Neural Networks*, 464 –471, 2002.

[28] S. Melacci and M. Belkin. Laplacian Support Vector Machines Trained in the Primal. *Journal of Machine Learning Research*, 12:1149–1184, March 2011.

[29] T. Mensink and J. J. Verbeek. Improving people search using query expansions. In *ECCV*, pages 86–99, 2008.

[30] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *ACCV, 2010.*, June 2008.

[31] D. Ozkan and P. Duygulu. A graph based approach for naming faces in news photos. In *CVPR*, 2006

[32] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[33] P. Pham, M.-F. Moens, and T. Tuytelaars. Naming persons in news video with label propagation. In *Proceedings of the International Workshop on Visual Content Identification and Search at the IEEE International Conference on Multimedia & Expo*, 1528–1533, 2010.

[34] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advance In Large Margin Classifiers*, 61–74, 1999.

[35] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Bipartite graph reinforcement model for web image annotation. In *ACM Multimedia*, 585–594, Augsburg, Germany, 2007.

[36] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3):157–173, 2008.

[37] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE MultiMedia*, 6(1):22–35, 1999.

[38] C. Siagian and L. Itti. Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29:300–312, 2007.

[39] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM Trans. Intell. Syst. Technol.*, 2:14:1–14:15, February 2011.

[40] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang. A face annotation framework with partial clustering and interactive labeling. In *CVPR*, 2007.

[41] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Image annotation refinement using random walk with restarts. In *Proceedings of ACM international conference on Multimedia*, pages 647–650, 2006.

[42] C. Wang, S. Yan, L. Zhang, and H.-J. Zhang. Multi-label sparse coding for automatic image annotation. *CVPR*, 0:1643–1650, 2009.

[43] D. Wang, S. C. Hoi, and Y. He. Mining weakly labeled web facial images for search-based face annotation. In *ACM SIGIR*, 535–544, 2011.

[44] D. Wang, S. C. H. Hoi, Y. He, and J. Zhu. Retrieval-based face annotation by weak label regularized local coordinate coding. In *ACM Multimedia*, pages 353–362, 2011.

[45] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: recognizing people and social relationships. In *ECCV*, pages 169–182, 2010.

[46] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image auto-annotation by search. In *CVPR*, 2006.

[47] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2), 2009.

[48] F. Wu, Y. Han, Q. Tian, and Y. Zhuang. Multi-label boosting for image annotation by structural grouping sparsity. In *ACM Multimedia*, pages 15–24. ACM, 2010.

[49] X. Wu and R. Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *ACM SIGKDD*, pages 326–333, New York, NY, USA, 2004.

[50] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum. Scalable face image retrieval with identity-based quantization and multi-reference re-ranking. In *CVPR*, pages 3469–3476, 2010.

[51] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. In *NIPS*, pages 2259–2267, 2009.

[52] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.

[53] J. Zhu. Semi-supervised learning literature survey. Technical report, Carnegie Mellon University, 2005.

[54] J. Zhu, S. C. Hoi, and L. V. Gool. Unsupervised face alignment by robust nonrigid mapping. In *ICCV*, 2009.

[55] J. Zhu, S. C. Hoi, and M. R. Lyu. Face annotation by transductive kernel fisher discriminant. *IEEE Transactions on Multimedia*, 10(01):86–96, 2008.

[56] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.